A PROPOSAL OF METRICS FOR DETAILED EVALUATION IN PRONUNCIATION MODELING

R. Barra, J. Macías-Guarasa, F. Fernández, L.F. D'Haro, J.M. Montero and J. Ferreiros

Grupo de Tecnología del Habla. Depart. de Ingeniería Electrónica. Universidad Politécnica de Madrid
E.T.S.I. Telecomunicación. Ciudad Universitaria s/n, 28040-Madrid, Spain
{barra, macias, efhes, lfdharo, juancho, jfl}@die.upm.es

ABSTRACT

In the context of ASR systems it is of major importance to accurately model the allophonic variations to be faced in a real world task.

The evaluation of which pronunciation variants are actually improving the system performance is crucial, as it determines the acceptance of the pronunciation alternatives used. Traditional approaches use different criteria and, typically, evaluation only cares about the global impact of the augmented dictionaries in the WER, so that this leads to little further insight on till what extent the proposed variations are actually working or not. Our proposal in this paper is also evaluating the *effective improvement* due to every pronunciation variation used, defining specific improvement metrics on the utterance level. We will show how these metrics actually highlight the beneficial impact achieved by the application of phonological rules when dealing with certain pronunciations variants, while the differences observed in global WER are not statistically significant.

1. INTRODUCTION

In the literature there are plenty of references to the problem of introducing pronunciation variants in speech recognition systems (an excellent revision can be found in [1]). Explicit modeling of the pronunciation alternatives is able to achieve substantial improvements if the acoustic models closely match the transcriptions [2]. The most usual strategies to add pronunciation variants use either knowledge based approaches, applying phonological rules to the canonical dictionaries, or data-derived pronunciation variants. The research teams are extremely careful when adding variants and several approaches to generate and limit their number have been proposed, such as using a ML criterion [3], smoothing the automatically derived phonetic transcriptions [4] or measuring the occurrences of variants [5], to name a few.

All of them try to evaluate till what extent the added variants are actually achieving better results or not. Unfortunately, traditional evaluation takes only into account the effectiveness of a given method in improving the overall WER. Detailed error analysis or alternative evaluation metrics do not seem to be a priority in order to provide insight into the processes underlying pronunciation variation (due to a number of practical reasons) [1].

In this scenario, when the overall WER evaluation leads to results that are not statistically significant, it is very difficult to decide whether the given pronunciation variants are effective or not, so that they may be easily thrown away without actually knowing whether their impact is relevant for a given subset of the database (for example, we may think of a minority set of speakers coming from an specific dialectal area).

In this paper, we propose a set of metrics and a methodology for its application, that can give additional insight on the detailed improvements that may be achieved by the application of phonological rules, when the differences observed in global WER are not statistically significant.

2. EXPERIMENTAL SETUP

In this paper, we will work in both isolated (IWR) and continuous speech recognition (CSR) systems using the hypothesis-verification paradigm, where we will consider pronunciation alternatives at the segmental level (more specifically, within word variations generated using manually-derived rules).

2.1. Speech recognition systems

In the IWR task [7], the hypothesis module follows a bottom-up approach in which a phonetic string build up algorithm (using CI semicontinuos HMMs) is followed by a lexical access stage. The verification module is based on the Viterbi algorithm, using context-dependent HMMs. The latter receives a list of *preselected* words (sorted according to their likelihoods) generated by the hypothesis module and generates the final recognition result.

In the CSR system [8], the hypothesis module uses an integrated search approach combining CD continuous

HMMS and bigram LM. It generates a word graph to be further rescored by the verification module. In the current version, the verification module uses additional information stored in a trigram language model.

In both the IWR and CSR tasks, we decided to use canonical pronunciations while training the acoustic models, as we want to assume the worst-case scenario regarding the trained HMMs.

2.2. Databases and dictionaries

In our IWR experiments, we have used a subset of the VESTEL database [9], composed of 9790 utterances. VESTEL is a realistic speaker-independent speech corpus collected over commercial telephone lines. Cross-validation is applied by means of a *leave-10%-out* strategy, in order to increase the statistical significance of the results. The *canonical dictionary* is composed of 1952 words, with a single phonetic transcription per word.

In our CSR experiments, we have used a subset of the INVOCA database that was designed to support research and development in spontaneous speech recognition systems in air traffic control tasks. INVOCA contains spontaneous conversations between air traffic controllers and airplane pilots in the Madrid-Barajas (MAD) airport [8]. Our results will be based on the evaluation of the *clearances subset*, composed of 8.9 hours of recorded conversations (5011 utterances), using 8 hours (4588 utterances) for training and 0.9 hours (503 utterances) for testing, with a *canonical dictionary* composed of 994 words (single phonetic transcription per word). The word graph is generated by an *n*-best search strategy during decoding (*n* between 10 and 20).

2.3. Rule selection process

From internal studies carried out in our Group, we generated an exhaustive repertoire of Castilian Spanish pronunciation variants, in order to be included in speech recognition systems. In this work, we further reduced this repertoire, leading to a selection of up to 13 phonological rules. The main selection criterion was to keep the rules commonly accepted as 'typical' for the average Spanish speaker. Nevertheless, our target is proposing and evaluating a set of performance metrics, so that the actual set of rules used is not that important. Rules are applied to the *canonical* dictionary in order to generate the *modified* one.

3. EVALUATION STRATEGY

In our proposal, the evaluation of pronunciation alternatives takes two factors into account:

- The (traditional) overall impact in WER, measured on the whole test set. It corresponds to a global, average performance, which we do not want to increase. However, we could tolerate a certain minimum degradation (i.e., statistically not significant), if this degradation leads to a benefit in some aspect that we may consider relevant.
- The (proposed) *effective improvement* of the considered rules, measured on the subset of the whole test set for which there exist differences between using or not the pronunciation alternatives (we call this improvement "*effective"* in the sense that it shows the actual detailed impact of the rules). This effect could seem irrelevant due to its limited impact in the overall performance (as pronunciation variants rarely have a big impact in global WER), but it is justified as it allows real-world systems to correctly recognize a certain set of speakers or pronunciations that, otherwise, would be poorly handled by the speech recognizer: in publicly deployed speech recognition systems, it is of outmost importance to reduce the number of speakers for which the system would not work at all (and allowing for a statistically not significant degradation in WER).

Our idea for designing new evaluation metrics has a common ground both in the IWR and CSR tasks: generating a specific *per-utterance performance metric* and integrating all the individual calculations in several quality metrics.

3.1. Evaluating *effective* improvements in the IWR task

The general idea in the IWR case is calculating in which position within the (sorted) list of recognized words, the correct word was actually recognized. So, the basic *per-utterance performance metric* would be the "rank" of the given utterance 'R': R_c when using the canonical dictionary and R_m when using the modified one. After making this calculation for both the canonical and modified dictionaries, we can compare the results and decide, for every utterance, which dictionary got the best result (the one with the lower R). The higher the differences between R_c and R_m, the higher the impact of the phonological rule applied.

We have evaluated a full set of up to 19 evaluation metrics, for the sake of brevity, we will just show a sample subset of them here:

- Number of utterances for which there was no difference between both dictionaries ($R_c = R_m$)
- For the subset utterances in which using the modified dictionary lead to better results ($R_c > R_m$). We will refer to this subset as 'improve' set:
  - Number of utterances
  - Average relative difference between R_m and R_c (measured as a percentage relative to R_c)

- For the subset of utterances in which using the modified dictionary lead to worse results ($R_c < R_m$). We will refer to this subset as 'worsen' set:
  - Number of utterances
  - Average relative difference (%) between R_c and R_m

Each of the proposed metrics (the ones shown here and the rest belonging to the full set of 19) gives more insight on specific aspects of the impact of the added variants (overall impact, absolute or relative improvement gain, details on canonical or modified dictionaries improvements and preferences, etc.) that would be impossible to evaluate by just looking at global WER.

3.2. Evaluating *effective* improvements in the CSR task

In the CSR task, the basic *per-utterance performance metric* has a natural candidate, the *per-utterance WER*, as proposed in [10], given its considerable improvement in informativeness and adequacy for statistical testing. When talking about *per-utterance WER*, we will refer to WER_c when using the canonical dictionary and WER_m when using the modified one.

From the full list of 15 evaluation metrics we have considered in [7] we just show a sample subset of them here:

- Number of utterances for which there was no WER difference between both dictionaries ($WER_c = WER_m$)
- For the 'improve' set: the subset of utterances in which the WER using the modified dictionary is better than when using the canonical one ($WER_c > WER_m$):
  - Number of utterances
  - Absolute and relative improvement of the average WER
- For the 'worsen' set: the subset of utterances in which the WER using the modified dictionary is worse than when using the canonical one ($WER_c < WER_m$):
  - Number of utterances
  - Absolute and relative improvement of the average WER

As in the IWR case, those metrics give us valuable detailed information, complementary to the global WER evaluation.

4. EXPERIMENTAL RESULTS

In order to validate our proposal and show how these metrics can be actually used and interpreted, we will give details on general evaluation metrics (search space increase and overall impact in error rate) and the proposed ones. The experimental procedure includes, as an example, the evaluation of a sample individual rule and the evolution of the results when applying the full set of rules, incrementally selected by means of the application of a *greedy* search algorithm (in which the search criteria is based on maximizing the error rate improvement for every addition of a new rule).

4.1. Evaluation in the VESTEL IWR task

The relative increase in dictionary size (compared to the canonical one) rises over 130% if we apply the full set of rules simultaneously. Relative increases when applying single rules vary from 0.14% and 37.60%. As an example, when applying the *dfinal* rule, the relative increase is around 2%.

When evaluating the overall impact in performance using any set of rules, we found that, while slightly increasing the error rate, the differences were not statistically significant. For example, when using the 13 phonological rules, the relative increase in inclusion error rate is below 0.06% (which is negligible, especially if we take into account that the number of entries in the dictionary has increased 130%). When only applying, e.g., the *dfinal* rule, we found a minor improvement in inclusion error rate: 1.38% relative.

In table 1 we show, as an example, the results for the application of the *dfinal* rule in the VESTEL IWR task.

Table 1: Effective improvement/worsening results when applying the *dfinal* rule (IWR task)

Metric	Value
Number of utterances for which *canonical = modified* ($R_c = R_m$)	9393 (96.6%)
Number of utterances for which *modified* is worse ($R_c < R_m$)	263 (2.7%)
Number of utterances for which *modified* is better ($R_c > R_m$)	64 (0.7%)
Avg. rel. worsening (R_c-R_m)/ R_m when *modified* is worse	**8.36%**
Avg. rel. improvement (R_m-R_c)/ R_c when *modified* is better	**70.25%**

The first important observation is that the number of words adversely affected by the added variants is higher than the number of benefited ones (263 vs. 64). From these figures, it is clear that the application of that rule will never have a statistically significant impact in overall performance and, if any, it will probably be negative. However, if we then focus on the average differences between R's, we can see that the improvement obtained with the modified dictionary is significantly larger than the improvements obtained with the canonical dictionary: every adversely affected word, in average, loses a 8.36% in relative position within the list of recognized words, while the benefited ones gain 70.25%, also in average. These figures show that the effective negative impact of

the application of the rule is much smaller than its effective positive impact in the words for which the recognition results are actually improved, with similar quantitative results for other rules.

Additional evaluation metrics show that the negative impact takes place for words recognized in high (bad) positions within the preselection list ($R_c \geq 100$), while positive impact takes place in low (good) positions ($R_m \approx 25$). As we are using a hypothesis-verification strategy and given the preselection list size, this means that, due to the introduction of pronunciation variants, the correct word will probably be included in the preselection list, thus having an opportunity to be correctly recognized by the verification module.

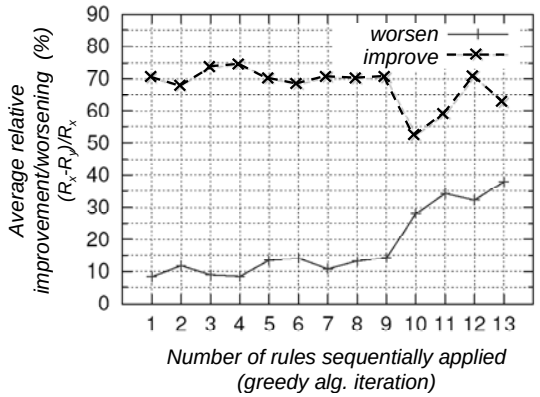

Figure 1: Average relative improvement/worsening in rank for the "worsen" and "improve" sets (IWR task)

Regarding the *greedy-based* incremental selection of optimal rules, Figure 1 shows the evolution of the average relative improvement in recognized position (rank) for the set of words benefited by the inclusion of pronunciation alternatives (*improve* in the figure) and the adversely affected ones (*worsen* in the figure), as a function of the number of rules incrementally applied (iterations of the *greedy* algorithm). We note that, for the first 9 best rules, the effective positive impact is much higher than the effective negative impact. This means that even though the number of utterances for which we can measure an improvement due to the use of pronunciation alternatives is low, the actual improvement obtained for such utterances is, again, much higher than the actual worsening due to the same alternatives. The main reason for such improvements is that the added pronunciation variants have a very strong positive effect in certain words that are poorly handled by the recognizer when using their canonical transcriptions, while leading to a minor impact in the rest of the words.

Plots such as the one shown in Figure 1 can give researchers a clue on how many rules should be used while achieving significant effective improvements due to the application of pronunciation alternatives (and provided that the overall impact in error rate is statistically not significant).

Table 2: Effective improvement/worsening results when applying the *dfinal* rule (CSR task)

Metric	Value
Number of utterances for which *canonical* = *modified* ($WER_c = WER_m$)	462 (91.9%)
Number of utterances for which *modified* is worse ($WER_c < WER_m$)	19 (3.8%)
Number of utterances for which *modified* is better ($WER_c > WER_m$)	22 (4.3%)
Avg. rel. WER worsening when *modified* is worse ($WER_c < WER_m$)	**31.46%**
Avg. rel. WER improvement when *modified* is better ($WER_c > WER_m$)	**53.33%**

4.2. Evaluation in the INVOCA CSR task

The relative increase in dictionary size if we apply the full set of rules rises over 173%. When applying single rules, relative increases vary from 0.56% to 32.24%. As an example, when applying the *dfinal* rule, the relative increase is 2.8%.

Again, when evaluating the overall impact measured as the increase in global WER, we obtained minor differences which were not statistically significant. For example, when using the 13 phonological rules, the relative increase in global WER is below 3% (negligible considering the 173% increment in number of entries). When only applying the, e.g., *dfinal* rule, the relative increase in global WER is below 0.3%. In Table 2 we show, as an example, the results for the application of the *dfinal* rule in the INVOCA CSR task.

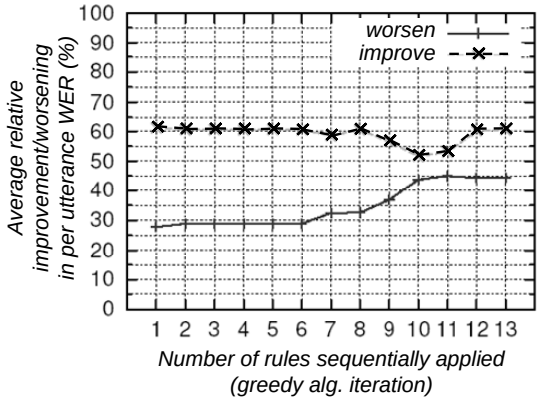

Figure 2: Avg. relative improvement/worsening in per-utterance WER (%) for the "worsen" and "improve" sets (CSR task)

In this case, the number of utterances adversely affected by the added variants is roughly the same than

the number of benefited ones (19 vs. 22), but, similarly to the IWR task, the average relative improvements due to pronunciation variants is higher than the performance loses (53.33% vs. 31.46%).

Figure 2 shows the evolution of the average relative improvement in WER (estimated from the word graph using an optimal search rescoring, and guided by the *greedy* algorithm searching for the optimal rules) for the *worsen* and *improve* set of utterances. Improvements due to pronunciation variants keep almost constant till the application of the 8th best rule and we could make similar considerations to the ones discussed in the IWR task: better inclusion results in the word graph lead to better results after rescoring.

4.3. Computational complexity considerations

As shown above, the increase in dictionary size can be high when using pronunciation variants. Our experiments show that, with the current systems using aggressive beam search techniques, the computational complexity increase due to increased dictionary size still allows for real time performance with the abovementioned results.

5. CONCLUSIONS AND FUTURE WORK

In this paper we have presented a proposal for the evaluation of pronunciation alternatives in speech recognition systems. It is aimed at complementing the traditional evaluation approach which only takes into account the impact in global error rate.

We have defined specific effective improvement metrics, especially adapted to get more insight on the actual impact of the pronunciation alternatives used, from different perspectives. All of the proposed metrics are based in the combination of a given *per-utterance quality metric* evaluated on the subsets for which differences between using or not pronunciation alternatives are found.

We have performed an experimental evaluation on both IWR and CSR tasks. Our results show that, even though the overall impact of the pronunciation alternatives is not statistically significant (even in dictionaries significantly bigger than the canonical ones), we can get effective benefits, important enough for certain users or pronunciations to have an opportunity to be correctly recognized. We show examples of the effect of including a single pronunciation rule and the incremental combination of all the selected rules by means of a *greedy* search algorithm. With the given metrics and the combination of evolution plots of these metrics, researchers can get additional information on till what extent individual rules are actually achieving effective improvements in the relevant database subsets.

To summarize, in our proposal the decision of including certain pronunciation variants is based on three considerations:

- The overall WER differences are not statistically significant
- The increase in computational load when using the variants is acceptable, given a certain hardware setup
- The relative improvements due to pronunciation variants is higher than the performance loses, according to the proposed metrics and plots

We are currently working in refining our methodology and specifically dealing with the evaluation of multiple pronunciation strategies using a data-driven approach.

6. ACKNOWLEDGEMENTS

This work has been partially supported under contracts TINA (UPM-CAM R05/10922), ROBINT (DPI2004-07908-C02) and EDECAN (TIN2005-08660-C04).

7. REFERENCES

[1] Strik, H. and Cucchiarini, C. “Modeling pronunciation variation for ASR: a survey of the literature”. Speech Communication, vol 29, p. 225-246. 1999.

[2] Saraçlar, M., Nock, H. and Khudanpur, S. “Pronunciation modeling by sharing Gaussian densities across phonetic models”. Computer Speech and Language, vol 14, p. 137-160. 2001.

[3] Holter. T. “Maximum Likelihood Modelling of Pronunciation in Automatic Speech Recognition”. PhD. Thesis. Norwegian University for S&T. 1998.

[4] Riley, M., Byrne, W et al. “Stochastic pronunciatio modeling from hand-labelled phonetic corpora”. Speech Communication, vol 29, p. 209-224. 1999.

[5] Kessens, J. and Wester, M. “Imrpving the performance of a Dutch CSR by modeling within-word and cross-word pronunciation variation”. Speech Communication, vol 29, p.193-207. 1999.

[6] Wester, M. "Pronunciation modeling for ASR – knowledge-based and data-derived methods". Computer Speech and Language, vol 17, p 69-85. 2003

[7] Macías-Guarasa, J., Ferreiros, J., Colás, J., Gallardo, A. and Pardo, J.M. “Improved Variable List Preselection List Length Estimation Using NNs in a Large Vocabulary Telephone Speech Recognition System”. ICLSP00, pp. 823-826. 2004

[8] Fernández, F., Córdoba, R., Ferreiros, J., Sama, V., D'Haro, L.F. and Macías-Guarasa, J. “Language Identification Techniques based on Full Recognition in an Air Traffic Control Task “ ICLSP04, pp. 1565-1568. 2004

[9] Tapias, D., Acero, A., Esteve, J., and Torrecilla, J.C. "The VESTEL Telephone Speech Database". ICSLP94, pp. 1811-1814. 1994.

[10] H. Strik, C. Cucchiarini, J.M. Kessens. “Comparing the performance of two CSRs: How to determine the significance level of the differences”. Eurospeech 2001, Vol. 3, pp. 2091-2094.